A Generalized Environmental Database for San Diego Bay

Gerald S. Key
Andrew E. Patterson
Marissa Caballero

Computer Sciences Corporation
4045 Hancock Street
San Diego, CA 92110-5164 USA
Internet mail: key@cscnet.com

INTRODUCTION

Understanding the fate and effect of materials introduced into marine ecosystems requires broad spatial and temporal perspectives. It may also require the aggregation of measurement data from different scientific disciplines. Because the collection of data on these scales is beyond the scope of most monitoring studies, it is often necessary to share measurements made by different investigators for different objectives. To do so requires the use of **primary measurement data recorded in fully documented digital form**.

A **primary measurement** is a quantitative observation made in the field or laboratory. It includes *what* was measured, the *quantity* of the measured parameter, and the *units* in which the quantity is expressed. Typical examples might include:

87.6 mg/kg lead
23 individual *Paralabrax clathratus*
3.6 cm/sec water velocity

Summary statistics such as means, standard deviations, and diversity indices are not suitable substitutes for primary measurement data. They account for less of the variance than the primary measurements and their methods of calculation are subject to change.

A **fully documented** primary measurement also includes supporting information that records *where* the measurement was made, *when* it was made, *how* it was made, *who* made the measurement, etc. These supporting attributes may have their own information requirements. For example, in reporting spatial position as latitude and longitude it is also necessary to report how the determination of position was made, the datum used as the coordinate reference, and the significant digits in the coordinate values. Fully documented measurements must also include associated measurements, such as those used to judge the quality of the data. Associated measurements might include measurements made on duplicates and replicate samples, the method detection limit, etc.

Environmental measurement data are typically voluminous; full documentation makes them more so. Measurement data must, therefore, be made available on a **digital** medium to facilitate storage, manipulation, and analysis of the data. Most environmental measurements made today are either recorded digitally or converted to digital form at some point in their life cycle. Reporting measurements digitally is in some sense easier than converting them to hardcopy. The reverse process of converting hardcopy records to a digital medium is too time-consuming, costly, and error-prone for most purposes.

This paper discusses efforts that are underway at the Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD) in San Diego, California to develop a generalized environmental data model. This model is being used to implement a multi-disciplinary environmental database, to guide the entry of historical data into the database, to prepare specifications for reporting fully documented measurements in future studies, and for sharing data with other projects with similar objectives.

BACKGROUND

The Environmental Sciences Division of NRaD has undertaken a study of San Diego Bay for the San Diego Naval Station (NavSta). The objective of this study is to understand the fate and effect of materials released into the Bay by the NavSta. In particular, the study is focusing on environmentally significant materials that have been associated with naval operations, including polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and heavy metals. The project staff is extracting measurements of these materials and related parameters from past studies and combining them with on-going Navy and non-Navy studies of the Bay. The data from all of these studies are being combined into a database that supports *ad hoc* query and reporting, direct interfaces to statistical, graphical, and other applications, and the extraction of data for use in simulation modeling and other external applications.

The design objectives for the NavSta data model are to provide:

Primary Measurements. The model should be able to represent all types of quantitative measurements made in the field and laboratory. The measurements should not be limited to a particular discipline (e.g., biology, chemistry), medium (e.g., sediment, water), or sample type (e.g., discrete, continuous).

Full Documentation. The data model should provide a template to ensure that all supporting and associated information about a measurement has been recorded. Multiple-use of environmental mea-surements, whether shared among contemporaneous projects or accumulated as a time-series for future studies, will invariably lead to unforeseen applications of the data. The re-use of measurements is limited if they lack the information required to judge the quality and applicability of the data to other uses.

No *a priori* View. Multiple-use of measurements also requires the data to be equally accessible to all perspectives. For example, storing data in a spatial data structure may be ideal for applications such as geographic information systems (GIS), but too restrictive for users interested in measurement methods, temporal distributions, etc. On the other hand, the data model, while not imposing a particular perspective on the data, should permit the user to reconstruct the perspective of the original investigation from the relationships to other data represented by the model.

Data Quality. The data model must provide the foundation for ensuring the quality and integrity of the data. It must rigorously define key fields, mandatory fields, and the types of relationships that exist between the data records. This information is essential in providing access and change control, security, and configuration management of the database.

Distributed Data. The data model should not preclude sharing data by linking databases at different locations via a network rather than combining the data in a central database.

Growth. The data model should accommodate new types of data without undue impact on the existing data structure or applications using those data structures.

A recent symposium at the University of New Mexico addressed the broad issue of environmental data management (Michener, *et al.*, 1994). Within this context, efforts are underway to develop national standards for representing environmental data. These efforts generally fall into two categories: metadata and geospatial data. **Metadata** are "data about data". They include information such as the objectives of the study that collected the data, the person to contact to obtain a copy of the data, where, when and how the data were collected, etc. Metadata are typically recorded in a separate companion file that is meant to accompany the measurement data file(s). In the United States, a draft standard for geospatial metadata has been developed by the Federal Geographic Data Committee (FGDC, 1994; see also http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html[1]).

As the name implies, **geospatial data** are data that identify the geographic location and characteristics of natural or constructed features and boundaries on the earth. Proposed content standards for geospatial data have their origin in the development of computer systems for storing and representing spatially distributed data, termed geographic information systems (GISs). The FGDC is leading the national effort to develop geospatial data standards, under the umbrella of the National Spatial Data Infrastructure (see http://fgdc.er.usgs.gov/ nsdi2.html). The Department of Defense is also taking an active role in geospatial data standards through the development of the Tri-Service Spatial Data Standards (see http://mr2.wes.army.mil/ docs/sds.htm). Finally, the Ecological Society of America has undertaken an effort to document and record long-term ecological data sets (see http://www.sdsc.edu/1/SDSC/Research/Comp_Bio/ESA/FLED/FLED.html).

The data model described in this report is complementary to these efforts. It incorporates many of the information attributes recorded in metadata files and links them directly to the corresponding measurement data. It also extends the scope of environmental data management to non-geospatial data.

RESULTS

Data Model. Figure 1 is an entity-relationship (E-R) diagram of the current NavSta data model. An E-R diagram is a logical representation of the *entities* about which data are to be stored, the *relationships* between those entities, and the *attributes* of the entities and their relationships[2].

[1] This and similar references are Uniform Resource Locators (URLs) to World-Wide Web (WWW) sites.

[2] An **entity** is a event, place, person, or concept about which we want to store information. In an E-R dia-

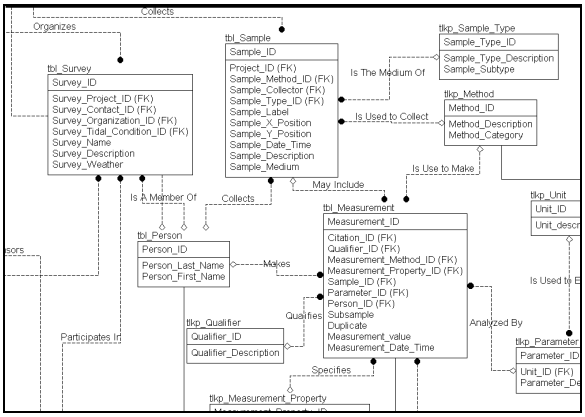

Figure 1. NavSta Data Model (part)

Entities can be viewed as two-dimensional tables, called *relations*, in which the columns (*attributes*) represent the pieces of information stored in each row (*record*) of the table. For instance, the relation **tbl_Measurement** in Figure 1 is designed to record information about measurements. It therefore includes attributes that identify what was measured (**Parameter_ID**), the quantity of the measured parameter (**Measurement_value**), the source of the measurement data (**Citation_ID**), etc. The attributes above the horizontal line in each relation in Figure 1 represent the *primary key* of the relation (e.g., **Measurement_ID**). The value (or combination of values) for the primary key uniquely identifies every record in that relation. Attributes designated with "(FK)" in Figure 1 are *foreign keys* - attributes that are nonkey (below the horizontal line) in one relation but part of the primary key in another relation.

Relationships link the primary key in one relation to an appropriate foreign key in another relation. Typically, relationships link the primary key value of a record in the "parent" relation to 0, 1 or many records in the "child" relation. The statement "a sample may include one or more measurements" is equivalent to saying that for each value of the primary key (**Sample_ID**) in the **tbl_Sample** relation there may be no, one, or many records of the **tbl_Measurement** relation with the same value in the **Sample_ID(FK)** attribute. This relationship is given the name **May Include** and it forms the basis for matching information about a sample with information about the measurements performed on that sample. Note that reading the **May Include** relationship in the other (Many:1) direction is also a true statement: "A measurement can belong to one and only one sample." Relationships may also be 1:1 and (rarely) Many:Many.

The complete NavSta model includes 29 relationships among 21 entities. Only the "core" entities and relationship, those directly related to representing measurements, are depicted in Figure 1. Both Figure 1 and the complete NavSta data model were generated using ERWin® by LogicWorks, Inc.

Database. The NavSta database has been implemented using Microsoft Corporation's Access® relational database management system (RDBMS; see McFadden & Hoffer, 1993) for Windows, in accordance with the specifications defined in the NavSta data model. Table 1 summarizes the size and composition of the NavSta database.

Number of:	
Measurements	30,031
Samples	733
Parameters	256
Data Sources	10
Media	2

Table 1: Summary Statistics for NavSta Database

While other members of the project staff were designing sampling programs that would generate new data for the NavSta database, the database staff set about identifying, acquiring, and entering data from previous studies in San Diego Bay. The staff prioritized these historical data sets according to measurement parameters, sample medium, proximity to the NavSta, and whether the data were available in digital form.

To date all of the primary measurement data (i.e., the parameter, quantity, units) that have been entered into the database were supplied in digital form, usually in a spreadsheet file. Most of the supporting information (e.g., longitude, latitude, sample date, etc.) and associated measurements (e.g., laboratory quality assurance values), when available, had to be key-entered from hardcopy documents. The average amount of time required to reorganize, merge, load, and validate these historical data set was 5 work-hours per 1,000 measurement records. Since these data sets were supplied in digital form, the time required to load them into the database was principally a function of the completeness and quality of the data, rather than the number of records.

---

gram, entities are represented as boxes. A **relationship** is an association (line in the E-R diagram) between two entities. **Attributes** are properties of the entity or the relationship.

The most common problems encountered converting these data sets have been:

Missing Data - no measurement units, methods; acronyms with no definition; longitude and latitude values with no datum

Units - converting comparable measurements to common units; converting position coordinates in State Planar, Marsden Squares, etc. to latitude and longitude.

Methods - deciding which methods are associated with which measurements.

Not Detected. It is common practice in environmental compliance studies to report a measured parameter as "ND" (Not Detected) when its value falls below the Method Detection Limit. Keith (1991) has discussed the arguments for and against this practice from the standpoint of analytical methods. From the perspective of database management, "ND" is a non-value. It cannot be stored in a numeric field nor can it be replaced by zero or the detection limit.

Obviously, many of the difficulties cited above could be avoided if environmental measurements were reported according to a pre-defined specification of the content and format of the data set. The NavSta project is preparing such a specification. Slagel (1994) has discussed the difficulty of developing data reporting standards.

The scope of the data entered into the NavSta database to date has been limited by the project's research objectives. Nonetheless, the staff has encountered a broad range of problems commonly associated with the design of environmental databases. One common problem is representing the relationship between hierarchically associated measurements, as illustrated in Figure 2.

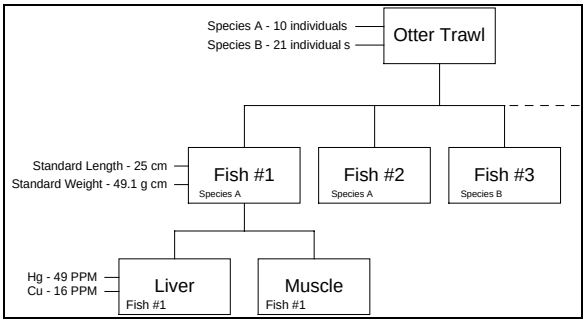

Figure 2. Sample Hierarchy

An objective of the database design is to make all measurements, whether population counts from an otter trawl or heavy metals from a particular tissue, equally accessible to the user. The user should be able to search for Parameter = "Cu" without having to know whether the measurement was made on a sediment, water, or tissue sample. However, having found a particular heavy metal measurement, the user should be able to retrieve the size and weight of the fish from which the liver was excised and a list of the other species that were caught in the same trawl - and what measurements were made on those organisms. Recursive associations such as this are usually maintained in an RDBMS in *unary* relationships. Unary relations are, however, difficult to maintain and to query.

A number of related issues arise from representing sample hierarchies. For example, sediment bioassays frequently measure the survival of test organisms that do not occur in the location where the sediment was collected. In effect, the test organisms become a "reagent" used to measure toxicity. This results in the scientific name of the test organism being relegated to the method description, where it is less useful as a query target. Bioassays are also representative of the problem of recording the measurement location vice the sample location. The fish counts used as examples in Figure 2 might have been made in the field, the lengths and weights migh have been made in the investigator's lab, and the tissue analyses performed weeks later and thousands of miles away at a contract laboratory. It is important to know where the measurements were made, but also to keep these locations separate from the locations where the measurements apply (i.e., the original sample site.)

CONCLUSIONS

The NavSta database has demonstrated that a multi-source, multi-disciplinary database can be developed in accordance with the specified design objectives. Such a database can be an effective tool for sharing data among members of the same project, and potentially between projects. The NavSta database effort has also underscored a number of areas requiring further research and development. Current and future efforts in this regard include:

Expanded Data Model. In conjunction with the Southwestern Division (SOUTHWESTDIV) of the Naval Facilities Engineering Command, efforts are underway to expand the NavSta data model to incorporate other dimensions of environmental data. In particular, the data model is being expanded to ac-

commodate the supporting data from hazardous waste studies and shore-based operations.

Distributed Database. In conjunction with the San Diego Supercomputer Center (SDSC), the NavSta project is investigating the use of client/server and distributed database architectures for managing environmental data. These investigations will entail both the horizontal (by record) and vertical (by attribute) distribution of the database, as well linking various client applications to the database server across the Internet. Among the distributed, client/server applications to be investigated are WWW browsers and GIS.

Data Reporting Specification. The NavSta project staff is working with SOUTHWESTDIV and SDSC, and through them with numerous other organizations, to develop a specification for reporting fully documented environmental measurement data. The objective is to use this specification for in-house data collection efforts and contracted studies to ensure the required data attributes are reported in a known format.

Object-Oriented Technology. Ultimately, the relational data model may be too limited for storing the complex data types and inter-relationships of environmental data. Object-oriented database (OODB) systems appear to address a number of these limitations and warrants closer scrutiny.

NRaD Database. The Environmental Sciences Division plans to expand to use of the NavSta database to be a long-term repository for a broader range of environmental studies at NRaD. NRaD is also studying the mechanisms and policy issues involved in making these data available to a regional data center such as the one being developed at the SDSC (see the paper by John Helly in this volume).

REFERENCES

FGDC. 1994. *Content standards for digital spatial metadata* (June 8 draft). Federal Geographic Data Committee. Washington, D.C.

Keith, L. H. 1991. *Environmental Sampling and Analysis: A Practical Guide*. Lewis Publishers, Chelsea, Michigan.

McFadden, F.R. and J.A. Hoffer. 1993. *Modern Database Management*, 4th edition. The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA.

Michener, W. K., J. W. Brunt and S. G. Stafford (eds). 1994. *Environmental Information Management and Analysis.* Taylor & Francis, London.

Slagel, R.L. (1994). "Standards for integration of multisource and cross-media environmental data", In: Michener, W. K., J. W. Brunt and S. G. Stafford (eds). 1994. *Environmental Information Management and Analysis.* Taylor & Francis, London.